CHEMICAL & PHARMACEUTICAL BULLETIN

Vol. 29, No. 10 October 1981

Regular Articles

Chem. Pharm. Bull.
29(10)2737—2742(1981)

A Least-squares Minimization Procedure to attain Optimum Spatial Fit between Homologous (Molecular) Structures and Its Application to Studying Protein Structural Homology

Yukihide Urata, Yukio Mitsui* and Kazuo Torii Nakamura

Faculty of Pharmaceutical Sciences, University of Tokyo,
Hongo, Tokyo 113, Japan

(Received February 18, 1981)

A theoretical basis and a program, LSFIT, for attaining optimum spatial fit between the two identical, similar or possibly related (molecular) structures (each described by an arbitrarily chosen coordinate system) are described. It is shown that the six relevant rotational and translational parameters relating the two coordinate systems can be derived by a standard iterative least-squares procedure for non-linear problems. Crystallographic applications and an application to the study of protein structural homology are described.

Keywords——protein structure; protein folding; structural homology; β-sheet topology; *Streptomyces* subtilisin inhibitor; immunoglobulin

Presenting the results of crystal structure analyses as computer-drawn pictures is a common practice. The well-known ORTEP system[1] has been widely used among crystallographers and other interested workers to prepare such pictures. Computer-drawn pictures provided as stereo-pairs are particularly useful to visualize the molecular structures and the vibrational mode of every atom. Other molecular graphics programs have been developed independently of ORTEP (*e.g.* the NAMOD system for general molecular structures (Beppu[2]) and the STDRAW system for proteins (Mitsui[3])). As crystal structure analyses have become increasingly easier, however, three-dimensional structures of many homologous or related compounds have become available and the need for comparing these structures visually has grown rapidly. To facilitate such visual comparison, however, it is vital to orient the set of molecules in question so as to optimize the fit between related structures. Mathematically, the optimum fit can be defined as that which makes the sum of the squares of the spatial displacements between the corresponding elements (usually atoms) in the related structures minimum, and the necessary procedure is to derive a set of rotational and translational parameters (RT parameters) giving best fit between the two relevant structures treated as rigid bodies. Deriving such RT parameters is not an obvious procedure. We devised a unique algorithm to achieve this and developed a FORTRAN program, LSFIT. As discussed later, the program turned out to be useful for many purposes other than that originally intended. While several other groups seem to have developed similar procedures, to our knowledge, they have never been described in full detail. In response to frequent requests for this procedure, we decided to describe in full the practical aspects of the procedure.

Theoretical

The Problem

Defining the three Eulerian angles (θ_1, θ_2, θ_3) in one of the conventional ways,[4,5] the positional vector $\boldsymbol{X}=(x_1, x_2, x_3)^t$ (t denotes transpose) generated from the original vector $\boldsymbol{X}'=(x_1', x_2', x_3')^t$ through translation vector $\boldsymbol{T}=(t_1, t_2, t_3)^t$ and rotation matrix R can be expressed as

$$\boldsymbol{X} = R\boldsymbol{X}' + \boldsymbol{T} \tag{1}$$

where

$$R = \begin{pmatrix} \cos\theta_1\ \cos\theta_3 & -\cos\theta_1\ \sin\theta_3 & \sin\theta_1\ \sin\theta_2 \\ \quad -\sin\theta_1\ \cos\theta_2\ \sin\theta_3 & \quad -\sin\theta_1\ \cos\theta_2\ \cos\theta_3 & \\ \sin\theta_1\ \cos\theta_3 & -\sin\theta_1\ \sin\theta_3 & -\cos\theta_1\ \sin\theta_2 \\ \quad +\cos\theta_1\ \cos\theta_2\ \sin\theta_3 & \quad +\cos\theta_1\ \cos\theta_2\ \cos\theta_3 & \\ \sin\theta_2\ \sin\theta_3 & \sin\theta_2\ \cos\theta_3 & \cos\theta_2 \end{pmatrix} \tag{2}$$

Suppose we have one rigid body and the coordinates of N representative points associated therewith ($\boldsymbol{X}_i'$ for the i-th representative point). In addition, we have the other rigid body having the same or closely related shape and size, and we know the coordinates of N representative points associated therewith ($\boldsymbol{X}_i$ for the i-th representative point which is supposed to correspond to the i-th point on the original rigid body). Then we should be able to derive the best R and $\boldsymbol{T}$ relating $\boldsymbol{X}'$ and $\boldsymbol{X}$. We have 6 independent unknowns (t_1, t_2, t_3, θ_1, θ_2, θ_3). Since the 9 elements of R are not linear with respect to θ_1, θ_2, θ_3, some means of deriving the initial values for the Eulerian angles to be subjected to an iterative least-squares refinement procedure has to be devised.

Derivation of the Initial Estimated Values

We temporarily treat the nine elements (r_{ij}, $i=1$—3, $j=1$—3) of R as if they were independent of each other, resulting in twelve "independent" variables (r_{ij}'s and t_i's). Then we select four sets (quartet) of ($\boldsymbol{X}_i'$ and $\boldsymbol{X}_i$) which are considered to be the best as regards both the accuracy of the coordinates and the correspondence within each set. Inserting these into equation (1), we get twelve linear equations, easily yielding the initial estimated values for t_i's, t_{10}, t_{20} and t_{30}. Since we have redundant data, the nine r_{ij}'s, the corresponding Eulerian angles, θ_{10}, θ_{20} and θ_{30}, can be derived in many ways including the following.

$$\begin{aligned} &\theta_2 = \cos^{-1} r_{33}\,(0\leqq\theta_2\leqq\pi) \\ &\theta_3 = \cos^{-1}(r_{32}/\sin\theta_2)(0\leqq\theta_3\leqq\pi) \\ &\qquad \text{but } \theta_3 = -\cos^{-1}(r_{32}/\sin\theta_2)(-\pi\leqq\theta_3\leqq 0),\ \text{if } r_{31}/\sin\theta_2<0 \\ &\theta_1 = \cos^{-1}(-r_{23}/\sin\theta_2)(0\leqq\theta_1\leqq\pi) \\ &\qquad \text{but } \theta_1 = -\cos^{-1}(-r_{23}/\sin\theta_2)(-\pi\leqq\theta_1\leqq 0),\ \text{if } r_{13}/\sin\theta_2<0 \end{aligned} \tag{3}$$

Each different derivation may yield somewhat different results due to the inherent errors in $\boldsymbol{X}$ and $\boldsymbol{X}'$. In the worst case, the process (3) may not work, because, for example, the absolute argument of arc cosine is significantly larger than unity. To cope with this difficulty, the program is designed to make trial calculations using various possible "quartet" sets.

Refinement of the Initial Estimated Values

This can be done by the usual iterative procedure for non-linear systems. Thus we define

$$S = \sum_{i=1}^{N} |\boldsymbol{X}_{i,\mathrm{obs}} - \boldsymbol{X}_{i,\mathrm{calc}}(\theta_{j0}+\Delta\theta_j(j=1\text{—}3), t_{j0}+\Delta t_j(j=1-3))|^2$$

for the N sets of representative points and minimize S with respect to $\Delta\theta_j$ and Δt_j.

For the three components (x_1, x_2, x_3) of $\boldsymbol{X}_{i,\mathrm{calc}}$ we have the expression,

$$\begin{aligned}(x_1)_i &= r_{11}(x_1')_i + r_{12}(x_2')_i + r_{13}(x_3')_i + t_1 \equiv f_{1i}(\theta_j,t_j(j=1\text{—}3))\\ (x_2)_i &= r_{21}(x_1')_i + r_{22}(x_2')_i + r_{23}(x_3')_i + t_2 \equiv f_{2i}(\theta_j,t_j(j=1\text{—}3)) \qquad (i=1\text{—}N)\\ (x_3)_i &= r_{31}(x_1')_i + r_{32}(x_2')_i + r_{33}(x_3')_i + t_3 \equiv f_{3i}(\theta_j,t_j(j=1\text{—}3))\end{aligned}$$

Here $(x_1)_i$, for example, can be expanded by means of the Taylor series into first-order terms of $\varDelta\theta$'s and $\varDelta t$'s,

$$\begin{aligned}(x_1)_i &\equiv f_{1i}(\theta_j,t_j(j=1\text{—}3)) \fallingdotseq\\ &f_{1i}(\theta_{j0},t_{j0}(j=1\text{—}3)) + \sum_{j=1}^{3}\{(\partial f_{1i}/\partial\theta_j)\varDelta\theta_j+(\partial f_{1i}/\partial t_j)\varDelta t_j\} \qquad (i=1\text{—}N)\end{aligned}$$

Thus we obtain $3N$ linear equations (the observational equations) with respect to $\varDelta\theta$'s and $\varDelta t$'s.

$$\begin{aligned}\sum_{j=1}^{3}\{(\partial f_{1i}/\partial\theta_j)\varDelta\theta_j+(\partial f_{1i}/\partial t_j)\varDelta t_j\} &= (x_1)_i - f_{1i}(\theta_{j0},t_{j0}(j=1\text{—}3))\\ \sum_{j=1}^{3}\{(\partial f_{2i}/\partial\theta_j)\varDelta\theta_j+(\partial f_{2i}/\partial t_j)\varDelta t_j\} &= (x_2)_i - f_{2i}(\theta_{j0},t_{j0}(j=1\text{—}3)) \qquad (i=1\text{—}N)\\ \sum_{j=1}^{3}\{(\partial f_{3i}/\partial\theta_j)\varDelta\theta_j+(\partial f_{3i}/\partial t_j)\varDelta t_j\} &= (x_3)_i - f_{3i}(\theta_{j0},t_{j0}(j=1\text{—}3))\end{aligned}$$

Using the notation $\overline{A}\big|^{m}_{6}$ denoting the matrix with m columns and n rows, this is rewritten as

$$\overline{A}\big|^{6}_{3N}\ \chi\big|_{6} = m\big|_{6N}$$

where $\boldsymbol{X}\big|_{6} = (\varDelta\theta_1,\ \varDelta\theta_2,\ \varDelta\theta_3,\ \varDelta t_1,\ \varDelta t_2,\ \varDelta t_3)^t$. Then the required normal equation is

$$\overline{A^t}\big|^{3N}_{6}\ \overline{A}\big|^{6}_{3N}\ \chi\big|_{6} = \overline{A^t}\big|^{3N}_{6}\ m\big|_{3N}$$

which reduces to

$$\overline{M}\big|^{6}_{6}\ \chi\big|_{6} = \varDelta\big|_{6}$$

where M, $\varDelta$ involve only the known quantities. From a practical point of view, an example of the components of the matrix A is given:

$$\begin{aligned}\partial f_{1i}/\partial\theta_1 &= (x_1')_i(-\sin\theta_1\cos\theta_3 - \cos\theta_1\cos\theta_2\sin\theta_3)\\ &\quad + (x_2')_i(\sin\theta_1\sin\theta_3 - \cos\theta_1\cos\theta_2\cos\theta_3) + (x_3')_i\cos\theta_1\sin\theta_2\\ \partial f_{1i}/\partial\theta_2 &= (x_1')_i\sin\theta_1\sin\theta_2\sin\theta_3 + (x_2')_i\sin\theta_1\sin\theta_2\cos\theta_3 + (x_3')_i\sin\theta_1\cos\theta_2\\ \partial f_{1i}/\partial\theta_3 &= (x_1')_i(-\cos\theta_1\sin\theta_3 - \sin\theta_1\cos\theta_2\cos\theta_3)\\ &\quad + (x_2')_i(-\cos\theta_1\cos\theta_3 + \sin\theta_1\cos\theta_2\sin\theta_3)\\ \partial f_{1i}/\partial t_1 &= 1;\ \partial f_{1i}/\partial t_2 = 0;\ \partial f_{1i}/\partial t_3 = 0\end{aligned}$$

Programming

Program LSFIT now consists of about 1000 FORTRAN statements and performs the following operations. 1) Read and store the coordinates of all the representative points on the "reference set" (the set described by vector $\boldsymbol{X}$) and the "moving set" (the set described above by vector $\boldsymbol{X}'$). 2) Pick up the four sets of representative points (the "quartet" described above) and derive the initial values for the three translational vectors and the three Eulerian angles using equation (3). Should the absolute value of the argument of arc cosine exceed unity, it is reduced to unity (with proper sign). The user can interrupt the calculation if the deviation from unity is too much, say more than 20%. 3) If the procedure 2) does not work well, the input coordinates and the choice of the "quartet" should be checked for possible errors or impropriety. Procedure 2) can be repeated with different quartet sets until satisfactory results are obtained. 4) Select N sets of corresponding representative points from the "reference set" and "moving set" to perform the least-squares refinement. At each end of the refinement cycle, the R-factor defined as $\sum_{i=1}^{N}|\boldsymbol{X}_{i,\text{obs}}-\boldsymbol{X}_{i,\text{calc}}|/\sum_{i=1}^{N}|\boldsymbol{X}_{i,\text{obs}}|$ and the deviation $|\boldsymbol{X}_{i,\text{obs}}-\boldsymbol{X}_{i,\text{calc}}|$ are printed for each set of representative points. The user can drop any "wrong" set as judged by these criteria either manually or automatically (by setting the maximum tolerable deviation in advance). 5) These procedures are repeated until the R-factor stops decreasing. The required number of cycles is normally 3 to 5 and the computing time is negligibly small. The program is available from the authors on request.

Results and Discussion

The present procedure was first devised as a means of correlating the RNase-S molecule (Y-molecule) packed in the Y-form crystal (which was solved at an atomic resolution)[6] with the chemically identical but crystallographically distinct RNase-S molecules (ZA-molecule and ZB-molecule) packed in the Z-form crystal as it appeared in the 4 Å resolution electron density map.[7] The three molecules were described with respect to arbitrarily chosen Cartesian orthogonal systems. Ten representative points were chosen on the surface of the ZA- and ZB-molecules making use of clearly defined protrusions or clefts (in the low-resolution electron density maps) which were easily identifiable with the corresponding atomic features in the Y-molecule. For the ten "representative points" thus chosen, the final root-mean-square displacements between the corresponding points were 0.59 Å and 0.72 Å for the ZA and ZB molecules, respectively.[7]

Other application examples follow.

1) The atomic coordinates read from Kendrew wire models of *Streptomyces* subtilisin inhibitor (SSI)[8] by the use of a semi-automatic coordinate measuring device[9] were fitted into the crystal lattice through the program LSFIT applied to five reference points. These reference points were first marked on the electron density map and then the corresponding markers were set within the frame of the wire model through the "half-mirror device,"[8] so as to give best fit. Conversion of the arbitrarily chosen coordinate system set on the model frame to the crystallographic fractional coordinate was thus made possible.

2) Conformational comparison between the hypothetical substrate of a microbial proteinase, subtilisin, and the corresponding reactive site of the macromolecular inhibitor SSI (Fig. 12 of Mitsui *et al.*[8]) and similar comparison between SSI and possibly related protein proteinase inhibitors (basic pancreatic trypsin inhibitor and soybean trypsin inhibitor) (Fig. 13 of Mitsui *et al.*[8]) have been successfully carried out through the use of the program LSFIT. The latter kind of comparison further led to the prediction of the possible mode of association between various proteinases and protein proteinase inhibitors, and led to a reasonable explanation for the origins of the proteinase specificities of protein proteinase inhibitors (Figs. 6 and 7 of Hirono *et al.*[10]).

3) The present technique was further extended to explore the possible structural and evolutional relationships between SSI and the variable region[11] of immunoglobulin (IMG). When the β-sheet topologies are expressed in a very schematic way after Levitt and Chothia,[12] it becomes clear that only a certain part of the β-sheet topology in the structure of IMG, among all the known three-dimensional structures, shows coincidence with that of SSI.[8] However, the comparison was so schematic, that we felt it necessary to compare the two structures in a more real sense as three-dimensional structures. As is clear from the above description, it is necessary to decide *a priori* which representative point on one object corresponds to which point on the other. Since the similarity between the five-stranded antiparallel β-sheets of IMG and SSI was only topological and there was no apparent correspondence between the amino acid sequences of the two proteins, deciding the corresponding sets of the α-carbons was necessarily more or less arbitrary, and the resulting root-mean-square displacement was as large as 13.9 Å initially. However, by drawing the α-carbon chains of both proteins at this stage (by use of

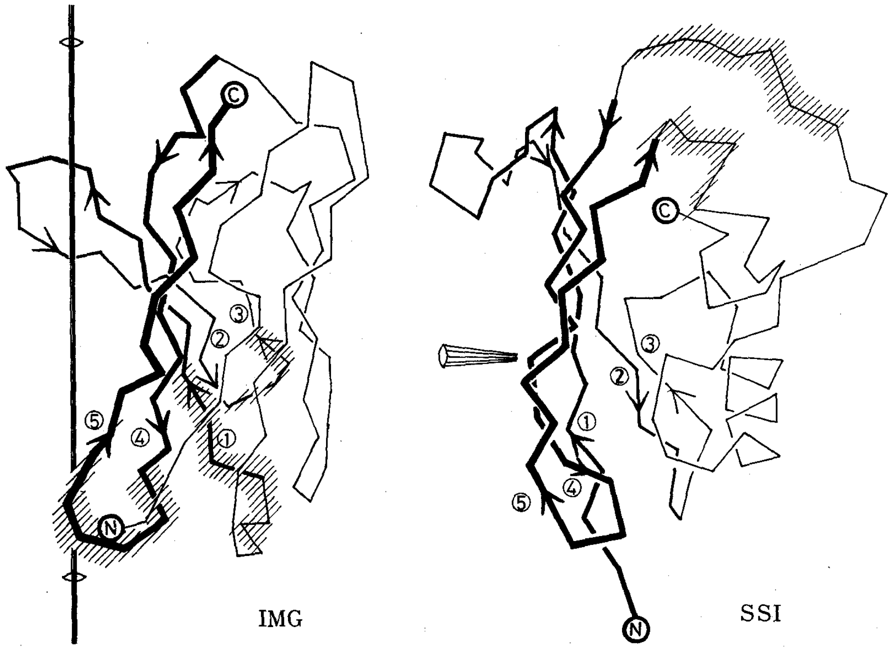

Fig. 1. The α-Carbon Chains of (a) Immunoglobulin Variable Region (IMG) and (b) One Subunit of *Streptomyces* Subtilisin Inhibitor (SSI)

The five apparently corresponding β-strands (numbered 1 through 5 differently stressed) were "optimally" superimposed as described in the text. The larger twist[14] in the five-stranded β-sheet of IMG compared to that of SSI is obvious. C- and N-termini are marked. No correspondences between the α-carbon chain foldings other than the five-stranded β-sheets are apparent between IMG and SSI. The functionally significant area (the hypervariable region[11] of IMG and the reactive site[8] of SSI) are shaded. The diads relating the shown subunit (in the case of SSI) or the domain (in the case of IMG) with the other are indicated at the left of each drawing. In the case of IMG, it runs vertically in the plane of the paper, while in the case of SSI, it makes an angle of 30° with the normal to the plane of the paper. The absence of geometrical correspondences of functionally significant regions and diads between IMG and SSI further suggests that the coincidence in the β-sheet topologies is accidental. See the text.

the program STDRAW[3] mentioned before), it was possible to select "better" sets of corresponding α-carbons by visual inspection. In the sceond run, the root-mean-square displacement was reduced to 5.7 Å for eight corresponding α-carbons. The drawing based on this result, after manual retouching to make the points of comparisons clearer, is shown in Fig. 1. As explained in the caption to Fig. 1, except for the topological resemblance of the two β-sheets, we see no other structural signs of an evolutional relationship between the two proteins. We thus support the general contention put forward by Richardson[13] that mere topological resemblance between the β-sheets of two different proteins should not be regarded as indicating an evolutional relationship unless other features of the protein molecules (if not the amino acid sequences) also indicate so.

The example 3) mentioned above highlights a basic weak point of the present procedure: the corresponding sets of representative points have to be decided by the user *a priori*. The decision is an obvious one when the two structures are almost identical (as in the case of example 1)) or closely related to each other (as in the case of example 2)). When the relationship is not obvious (as in the case of example 3)), however, the decision cannot be objective and the result (the "optimum fit") necessarily reflects the intention of the user. To obtain a truly objective optimum fit between two such structures, some other program based on completely different algorithm, such as the so-called vector superposition method in crystallography, would be necessary.

References and Notes

1) C.K. Johnson, Oak Ridge Temperature Ellipsoid Program ORTEP (1962).

2) Y. Beppu, Computer Center of the University of Tokyo, Center News, **11**, Suppl. 2, 64 (1978).
3) Y. Mitsui, Y. Watanabe, Y. Urata, and K.T. Nakamura, Abstracts of the 6th Symposium on Biomolecular Structure (Osaka), 1979, p. 15.
4) M.G. Rossmann and D.M. Blow, *Acta Crystallogr.*, **15**, 24 (1962).
5) H. Margenau and G.M. Murphy, "The Mathematics of Physics and Chemistry," D. Van Nostrand, Princeton (1943).
6) H.W. Wyckoff, D. Tsernoglou, A.W. Hanson, J.R. Knox, B. Lee, and F.M. Richards, *J. Biol. Chem.*, **245**, 305 (1970).
7) K. Torii, Y. Urata, Y. Iitaka, F. Sawada, and Y. Mitsui, *J. Biochem.*, **83**, 1239 (1978).
8) Y. Mitsui, Y. Satow, Y. Watanabe, and Y. Iitaka, *J. Mol. Biol.*, **131**, 197 (1979).
9) Y. Mitsui, *J. Appl. Crystallogr.*, **12**, 135 (1979).
10) S. Hirono, K.T. Nakamura, Y. Iitaka, and Y. Mitsui, *J. Mol. Biol.*, **131**, 855 (1979).
11) O. Epp, P. Colman, H. Fehlhammer, W. Bode, M. Schiffer, and R. Huber, *Eur. J. Biochem.*, **45**, 513 (1974).
12) M. Levitt and C. Chothia, *Nature* (London), **261**, 552 (1976).
13) J.S. Richardson, *Nature* (London) **268**, 495 (1977).
14) C. Chothia, *J. Mol. Biol.*, **75**, 295 (1973).